BIOINFORMATICS INSTITUTE

Searching for DNA rescue proteins in tardigrades

Daria Vashunina[1,2] **and Aleksei Osipov**[1,3]

[1]Bioinformatics Institute, Saint Petersburg, Russia
[2]Yandex, Moscow, Russia
[3]ITMO University, Russia

Abstract

In this study, we investigate the genetic and physiological basis of Tardigrade outstanding stability under harmful conditions on genome of *R. varieornatus*. Structural annotation was performed on the assembled genome excluding repeats and matched to the peptides obtained from the nuclei by Mass Spectrometry. After that, DNA-associated proteins were again sorted using TargetP and WoLF leaving us six proteins that were not recognized by blastp including one damage suppressor protein renowned today as Dsup. Latter investigation of these 6 genes, especially Dsup may open new horizons in understanding of DNA-damage resistance.

Keywords: Tardigrades, damage suppressor protein, subcellular localization, functional annotation.

Introduction

Tardigrades, also called water bears, are a unique type of multicellular model animals well known for their ability to survive in extreme conditions, including dehydration, freezing, high salinity, altitude and even radiation and open space. They are spreaded all around the world from salty lakes to the highest peaks sharing one common property required for active metabolism - presence of surrounding water. At the same time these small life beings are capable of cryptobiosis, a reversible ametabolic state possessed by desiccation and bioprotectants synthesis in order to withstand the hardest conditions on Earth and beyond, but they also capable of adaptation retaining active state[1]. In this case DNA repairing mechanisms come to the forefront what is really interesting in terms of medicine, biotechnology and ecology[2].

Using BATI algorithm (Blast, Annotate, Tune, Iterate), which utilizes Tblastn results to annotate intron/exon boundaries of genes ab initio and SWISS-MODEL to evaluate potential function, Carrero et. al have shown multiplication of MRE11 or XPC, and missense mutations such as CHEK1, POLK, UNG and TERT[3]. Some studies focus on comparative genomics and transcriptomics in close species *R. varieornatus* and *H. dujardini*. The 2 species differ in their responses to anhydrobiosis which in turn could be affected by horizontal gene transfer demonstrated by the HGT index approach. Protein family analyses and comparative genomics performed with Mauve, blastx, Ensembl Perl API demonstrated single-copy orthologues between *H. dujardini* and *R. varieornatus* using the orthologous groups defined by OrthoFinder[4].

Our study is going to be focused on the assumption about the nuclei localization of reparation proteins in *R. varieornatus*. In order to annotate and identify these proteins, we need to perform the repeat masking with RepeatModeller[5] and get rid of misleading intron regions that can mess up the annotation. For gene prediction the main way "de novo" is GeneMark-ES[6] based on the Markov models requiring *Viterbi-type algorithm*. Another way opens up when we get hint represented as a transcriptome and we can use Augustus[7] to make such predictions. This work provides insights into the hidden survival mechanisms of invincible water bear, which can help us to understand how to maintain our DNA safe and sound.

Methods

Obtaining data

For this project we used preassembled genome of *Ramazzottius varieornatus*, the YOKOZUNA-1 strain[8]. The genome have been downloaded from NCBI database in .fasta format. Regions of interest in the genome of tardigrades, corresponding to genes and proteins, were obtained using AUGUSTUS[9]. We downloaded precomputed results kindly provided by the course instructors: protein fasta and gff. The list of DNA-associated peptides from experimental data (tandem mass spectrometry) also

has been provided.

DNA-assosiated proteins search

To obtain DNA-associated proteins we did a local alignment-based search. We created a local database from our protein fasta file and run blastp (BLAST[10] 2.16.0+ Protein-Protein) on the experimental peptide sequence file as a query with default settings. Exact commands are provided in our lab journal.

Localization prediction

WoLF PSORT[11] was used for proteins localization prediction, organism type was selected as Animal. Also we used TargetP 2.0 for subcellular localization of the proteins, organism type was selected as Non-plant. BLAST web-version was used for proteins search against UniProtKB/Swiss-Prot database.

Function prediction

hmmscan tool from HMMER web-version was used for protein function prediction based on similar domains and motifs.

Results

The total amount of proteins in precomputed augustus.whole.aa file was 16435. After we run blastp on our experimental peptides list we got 34 DNA-associated proteins. You can find the file with all the 34 proteins in the supplementary materials. After running WoLF on those 34 proteins, we got different types of localization: nucleus, plasma membrane, cytosol and extracellular. TargetP found several proteins with secretory pathway signal peptide sequence (SP) and the rest with OTHER localization type. Pfam found known functional domains for 21 of 34 proteins. After running BLAST on all 34 proteins the we got some matches with sequences in other organisms. For those matches BLAST accession numbers, e-value, identity and query coverage persentages are represented in Table 1. in Supplementary Materials as well as WoLF, Target and Pfam results.

34 proteins is still too much for experimental search of the one or several related to DNA-defense mechanism. We performed filtering of the proteins based on their localization and function prediction result. From WoLF results we need only those ones with nuclear localization and from TARGET with OTHER localization. As we are looking for something definitely new and unknown, we filtered out all the proteins that match with known proteins in BLAST and functional domains in Pfam. As a result of filtering we got 6 proteins represented in Table 2. We suggest all of them as candidates for tardigrade-unique DNA-associated protein related with DNA defense functions.

Protein	WOLF	TARGET	BLAST, Accession Number	E-value	% Ident	% Query cov	Pfam, ID
g10513.t1	nucl	OTHER	-	-	-	-	-
g10514.t1	nucl	OTHER	-	-	-	-	-
g11806.t1	nucl	OTHER	-	-	-	-	-
g14472.t1	nucl	OTHER	-	-	-	-	-
g16318.t1	nucl	OTHER	-	-	-	-	-
g16368.t1	nucl	OTHER	-	-	-	-	-

Table 2. DNA-defending candidate proteins.

Discussion

We compared our results with the previous work of Takuma Hashimoto et al.[12] They identified a new tardigrade-specific DNA-associated damage suppressor protein and provide its peptide sequence: LTSSGTGAGSAPAAAK. We have searched through our 6 potential target proteins and got the exact match for g14472.t1. So we can make a conclusion that g14472.t1 is a DNA-associated damage suppressor protein. Our proteins g10513.t1 and g10514.t1 also have been found in Takuma Hashimoto et al. work as candidate proteins (RvY 12593 and RvY 12594). As in case with Dsup, we compared their peptide sequences with our 6 target proteins. All of these proteins may be responsible for different sustainability circuits since there are several ways to damage and restore DNA strands. It can be both Single strand breaks(SSB) or Double strand breaks(DSB). Radiation usually cause severe damage for nucleotide strands, so DSBs take place in this case, recalling Non-Homologous End Joining, main way for G1 phase, Homologous End Joining(HEJ) active in S and G2 phases and Micro-homologous End Joining(MHEJ) an alternative way for all of these stages. In order to figure out how these proteins act it is reasonable to look at the cell cycle phase distribution on modified cells under stress and try to evaluate expression of different cells (that's great if you have money to do scRNA, but usually u don't (:, ha-ha, but we still can perform it using Western Blot or RT-PCR at least). It is also crucial to look at the well known repair proteins as they are the best pathfinders on the DDR road (p21, ATR, ATM, CHK1. CHK2). Pathways that are active in cancerous resistant cells similar to JAK/STAT would be also great candidates to investigate.

Supplementary Materials

34 proteins fasta
Full summary results table

Protein	WoLF	TARGET	BLAST, Accession	E-value	% Ident	% Query cov	Pfam, ID
g10513.t1	nucl	OTHER	-	-	-	-	-
g10514.t1	nucl	OTHER	-	-	-	-	-
g11320.t1	plas	SP	-	-	-	-	-
g11513.t1	cyto	OTHER	Q32PH0.1	7E-83	29%	68%	TRAPPC9-Trs120
g11806.t1	nucl	OTHER	-	-	-	-	-
g11960.t1	nucl	OTHER	Q8CJB9.1	6E-98	27%	96%	zf-C3HC4
g12388.t1	extr	SP	P0DPW4.1	3E-11	38%	50%	CBM_14
g12510.t1	plas	OTHER	-	-	-	-	-
g12562.t1	extr	SP	P0DPW4.1	7E-13	40%	41%	CBM_14
g1285.t1	extr	SP	P0DPW4.1	2E-12	37%	44%	CBM_14
g13530.t1	extr	SP	-	-	-	-	-
g14472.t1	nucl	OTHER	-	-	-	-	-
g15153.t1	extr	SP	P0DPW4.1	2E-14	40%	46%	CBM_14
g15484.t1	nucl	OTHER	Q155U0.1	0	45%	78%	VPS51_Exo84_N
g16318.t1	nucl	OTHER	-	-	-	-	-
g16368.t1	nucl	OTHER	-	-	-	-	-
g2203.t1	plas	OTHER	Q69ZQ1.2	2E-126	36%	75%	Glyco_hydro_31_2nd
g3428.t1	mito	OTHER	Q09510.1	9E-65	57%	91%	EF-hand_7
g3679.t1	extr	SP	Q19269.2	7E-22	30%	72%	Astacin
g4106.t1	E.R.	OTHER	-	-	-	-	-
g4970.t1	plas	OTHER	Q9JIQ8.3	1E-15	26%	43%	Trypsin
g5237.t1	plas	OTHER	-	-	-	-	-
g5443.t1	extr	OTHER	-	-	-	-	-
g5467.t1	extr	SP	P0DPW4.1	4E-13	44%	46%	CBM_14
g5502.t1	extr	SP	P0DPW4.1	6E-14	40%	33%	CBM_14
g5503.t1	extr	SP	P0DPW4.1	7E-14	40%	38%	CBM_14
g5510.t1	plas	OTHER	-	-	-	-	MARVEL
g5616.t1	extr	SP	P0DPW4.1	2E-14	41%	40%	CBM_14
g5641.t1	extr	SP	P0DPW4.1	5E-13	39%	43%	CBM_14
g5927.t1	nucl	OTHER	Q17427.1	1E-18	39%	14%	-
g702.t1	extr	SP	P0DPW4.1	1E-11	40%	39%	CBM_14
g7861.t1	nucl	OTHER	B4F769.1	2E-71	37%	99%	SNF2-rel_dom
g8100.t1	nucl	OTHER	Q2YDR3.1	3E-46	36%	22%	Inositol_P
g8312.t1	nucl	OTHER	Q5KU39.1	0	41%	84%	Clathrin

Table 1. Summary results for proteins localization and function.

References

1. Møbjerg N., N. R. C. New insights into survival strategies of tardigrades. *Comp. Biochem. Physiol. Part A: Mol. Integr. Physiol.* **254**, 110890 (2021).
2. Jönsson, K. I. Radiation tolerance in tardigrades: current knowledge and potential applications in medicine. *Cancers* **11** (2019).
3. Carrero D., e. a. Differential mechanisms of tolerance to extreme environmental conditions in tardigrades. *Sci. reports* **9** (2019).
4. et al., Y. Y. Comparative genomics of the tardigrades hypsibius dujardini and ramazzottius varieornatus. *PLoS biology* **15** (2017).
5. et al., F. J. M. Repeatmodeler2 for automated genomic discovery of transposable element families. *Proc. Natl. Acad. Sci.* **117**, 9451–9457 (2020).
6. Brůna T., B. M., Lomsadze A. Genemark-ep+: eukaryotic gene prediction with self-training in the space of genes and proteins. *NAR genomics bioinformatics* (2020).
7. Stanke M., M. B. Augustus: a web server for gene prediction in eukaryotes that allows user-defined constraints. *Nucleic acids research* (2005).
8. Horikawa, D. D. The tardigrade ramazzottius varieornatus as a model animal for astrobiological studies. *Astrobiology* (2008).
9. Stanke M., M. B. Augustus: a web server for gene prediction in eukaryotes that allows user-defined constraints. *Nucleic acids research* (2005).
10. Camacho. Blast+: architecture and applications. *BMC Bioinforma.* (2009).
11. Horton, P. Wolf psort: protein localization predictor. *Nucleic acids research* (2007).
12. Hashimoto, T. Extremotolerant tardigrade genome and improved radiotolerance of human cultured cells by tardigrade-unique protein. *Nat. communications* (2016).